DOCUMENT RESUME

ED 138 633 TM 006 260

AUTHOR Hollingsworth, Holly
TITLE On the Comparison of Two Dependent Multiple Correlations.
PUB DATE [Apr 77]
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Comparative Statistics; *Correlation; Matrices; Predictor Variables
IDENTIFIERS Confidence Intervals (Statistics); *Multiple Correlation

ABSTRACT

A theorem of Spjotvoll (1972) was used to determine and apply a confidence interval for the difference of two multiple correlations based on observations from a single sample. Spjotvoll's method of comparing regression functions is also applicable to a comparison of dependent multiple correlations, an unsolved problem posed by Hotelling in 1940. The computations necessary to make the comparisons are explained. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

22.14

ED138633

On the Comparison of Two Dependent Multiple Correlations*

Holly Hollingsworth
Graduate School of Education
University of Pennsylvania

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

1. Introduction

Rudi Spjøtvoll (1972) developed a theorem which is applicable, under certain conditions, to the problem of finding simultaneous confidence intervals about a quadratic form, $\underset{\sim}{\phi}'\underset{\sim}{C}\,\underset{\sim}{\phi}$. The subject of Spjøtvoll's paper was the comparison of two or more regression functions derived from the same sample. In practice, these different regression functions arise from different predictor subset selection procedures, e.g., stepwise or backward elimination.

The purpose of this paper is to show that Spjøtvoll's method of comparing regression functions, when taken a step further, is applicable to comparing dependent multiple correlations. The problem of comparing dependent multiple correlations is, to date, an unsolved problem posed by Harold Hotelling in 1940. Hotelling (1940) showed that because of the presence of nuisance parameters, the difference of two dependent squared multiple correlations does not have an "exact or 'studentized' distribution" (p. 282). The solution to be offered here will be a conservative solution in the same sense that the Scheffé method of multiple comparisons is conservative. In point of fact, the method of Spjøtvoll is based on the same theoretical considerations that the S-method of multiple comparisons were based.

A simultaneous confidence interval will be constructed about the difference of two squared multiple correlations and if the interval does not contain the point zero, we may reject the hypothesis of equal multiple correlations. It is believed that because the intervals are simultaneous confidence intervals based on the Scheffé method of multiple comparisons

TM006 260

*Paper presented at the Annual Meeting of the American Educational Research Association, April, 1977.

that the length of the interval will be conservatively long. Future research in this area may want to investigate the applicability of the theory of Tukey's method of multiple comparisons to reduce the length of the confidence interval.

This paper will also indicate the necessary computations to make comparisons of dependent multiple corr[illegible]ns using the results of Spjøtvoll. The method will be demonstrat[illegible] on a realistic data set.

2. Method

Let $\hat{\underset{\sim}{\psi}}$ be a q-component random vector with distribution $N(\underset{\sim}{\psi},\sigma^2\underset{\sim}{B})$ where $\underset{\sim}{\psi}$ and σ are unknown parameters, and $\underset{\sim}{B}$ is a known positive definite matrix. Let s^2 be an estimate of σ^2 independent of $\hat{\underset{\sim}{\psi}}$ such that $\nu s^2/\sigma^2$ has a Chi-square distribution with ν degrees of freedom.

Let $\underset{\sim}{C}$ be a given qxq symmetric matrix and let $\underset{\sim}{P}$ be a qxq non-singular matrix such that $\underset{\sim}{P}'\underset{\sim}{B}^{-1}\underset{\sim}{P} = \underset{\sim}{I}$ and $\underset{\sim}{P}'\underset{\sim}{C}\underset{\sim}{P} = \underset{\sim}{D}$ where $\underset{\sim}{D}$ is a diagonal matrix with real diagonal elements $d_1,d_2,\ldots,d_q$ equal to the roots of $|\underset{\sim}{C} - d\underset{\sim}{B}^{-1}| = 0$.

Let $\hat{\underset{\sim}{\gamma}} = \underset{\sim}{P}^{-1}\hat{\underset{\sim}{\psi}}$ and let λ_{min} and λ_{max} be the minimum and maximum roots, respectively, of

$$\sum_{i=1}^{q} d_i^2\hat{\gamma}_i^2/(d_i - \lambda)^2 = c, \quad (\text{ or } \hat{\underset{\sim}{\gamma}}'\underset{\sim}{D}(\underset{\sim}{D} - \lambda\underset{\sim}{I})^{-2}\underset{\sim}{D}\hat{\underset{\sim}{\gamma}} = c)$$

where $c = qs^2F(1-\alpha:q,\nu)$ is a multiple of an F-tabled value. Finally, let $a = -\min(\min d_i, \lambda_{min})$ and $b = \max(\max d_i, \lambda_{max})$.

The following result, stated without proof, is due to Spjøtvoll.

THEOREM With the above assumptions and notation, the probability is at least $1 - \alpha$ that simultaneously for all quadratic functions $\underset{\sim}{\psi}'\underset{\sim}{C}\underset{\sim}{\psi}$ we have

$$A_1 \leq \underset{\sim}{\psi}'\underset{\sim}{C}\underset{\sim}{\psi} \leq A_2$$

where $A_1 = a\left(\sum_{i=1}^{q} d_i\hat{\gamma}_i^2/(a + d_i) - c\right)$ $\qquad$ $\left(\text{or } A_1 = a(\hat{\underset{\sim}{\gamma}}'\underset{\sim}{D}(a\underset{\sim}{I} + \underset{\sim}{D})^{-1}\hat{\underset{\sim}{\gamma}} - c)\right)$

and $A_2 = b(\sum_{i=1}^{q} d_i\hat{\gamma}_i^2/(b - d_i) + c)$ (or $A_2 = b(\hat{\underset{\sim}{\gamma}}'\underset{\sim}{D}(b\underset{\sim}{I} - \underset{\sim}{D})^{-1}\hat{\underset{\sim}{\gamma}} + c)$)

with exception of the following cases:

$A_1 = 0$ if all $d_i \geq 0$ and $\sum_{i\in I} \hat{\gamma}_i^2 \leq c$

$A_2 = 0$ if all $d_i \leq 0$ and $\sum_{i\in I} \hat{\gamma}^2 \leq c$

where $I = \{i: d_i \neq 0\}$.

3. Comparison of Dependent Squared Multiple Correlations

Suppose we have the regression model $\underset{\sim}{y} = \underset{\sim}{X}\beta + \underset{\sim}{e}$ where $\underset{\sim}{X}$ is a nxq, full column rank matrix, $\underset{\sim}{y}$ is a nx1 vector and the n-component random vector $\underset{\sim}{e}$ is distributed as $N(\underset{\sim}{0},\sigma^2\underset{\sim}{I})$. Given two matrices $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$, where $C(\underset{\sim}{X}_i)\subset C(\underset{\sim}{X})$, i = 1 or 2, ($C(\underset{\sim}{A})$ denotes the space spanned by the columns of the matrix $\underset{\sim}{A}$), we wish to compare the squared multiple correlations of $\underset{\sim}{y}$ with $\underset{\sim}{X}_1\underset{\sim}{\beta}_1$, R_1^2, and $\underset{\sim}{y}$ with $\underset{\sim}{X}_2\underset{\sim}{\beta}_2$, R_2^2, where it is assumed that there exists vectors $\underset{\sim}{\beta}_1$ and $\underset{\sim}{\beta}_2$ such that $E(\underset{\sim}{y}) = \underset{\sim}{X}_i\underset{\sim}{\beta}_i$, i = 1 or 2.

Note that $R_i^2 = \underset{\sim}{\beta}_i'\underset{\sim}{X}_i'\underset{\sim}{X}_i\underset{\sim}{\beta}_i/\sigma^2$, i = 1 or 2, so that $R_1^2 - R_2^2 = 0$ if and only if $\underset{\sim}{\beta}_1'\underset{\sim}{X}_1'\underset{\sim}{X}_1\underset{\sim}{\beta}_1 - \underset{\sim}{\beta}_2'\underset{\sim}{X}_2'\underset{\sim}{X}_2'\underset{\sim}{\beta}_2 = 0$.

Now then, since the least squares estimate of $\underset{\sim}{\beta}_i$ is $\hat{\underset{\sim}{\beta}}_i = (\underset{\sim}{X}_i'\underset{\sim}{X}_i)^{-1}\underset{\sim}{X}_i'\underset{\sim}{y}$ we have that $\underset{\sim}{\beta}_i = E(\hat{\underset{\sim}{\beta}}_i) = (\underset{\sim}{X}_i'\underset{\sim}{X}_i)^{-1}\underset{\sim}{X}_i'\underset{\sim}{X}\underset{\sim}{\beta}$, i = 1 or 2.

Thus $\underset{\sim}{\beta}_1'\underset{\sim}{X}_1'\underset{\sim}{X}_1\underset{\sim}{\beta}_1 - \underset{\sim}{\beta}_2'\underset{\sim}{X}_2'\underset{\sim}{X}_2\underset{\sim}{\beta}_2 = \underset{\sim}{\beta}'(\underset{\sim}{X}'\underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'\underset{\sim}{X} - \underset{\sim}{X}'\underset{\sim}{X}_2(\underset{\sim}{X}_2'\underset{\sim}{X}_2)^{-1}\underset{\sim}{X}_2'\underset{\sim}{X})\underset{\sim}{\beta}$

and the quadratic form on the right hand side is exactly the difference of goodness between the two regression functions as defined by Spjøtvoll, p. 1082, eq. 5.2.

4. Computational Steps

The following steps indicate the computations necessary to determine the confidence intervals. Steps A through G are given in matrix notation and can be easily completed using matrix subroutines offered in many software

packages. Step H involves the determination of the roots of a polynomial equation and intervals are given for those roots. If computer software is not available for this, the methods of Newton-Raphson or Horner are applicable and are discussed in most theory of equation texts.

A. Transform the predictor and criterion scores to scores having mean zero and variance $1/(N-1)$. Call the matrix of transformed predictors $\underset{\sim}{X}$ and the vector of transformed criterion scores $\underset{\sim}{y}$.

B. Compute $\underset{\sim}{X}'\underset{\sim}{X}$ and $\underset{\sim}{y}'\underset{\sim}{X}$, noting that these matrices are the intercorrelations of the predictors and the correlations of the predictors with the criterion, respectively.

C. i) compute $\underset{\sim}{\beta}'$ as $\underset{\sim}{y}'\underset{\sim}{X}\,(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$
ii) compute $s^2 = \underset{\sim}{y}'\,(I - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}')\,\underset{\sim}{y}/n-q$
iii) determine the partitions $\underset{\sim}{X}_1'\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2'\underset{\sim}{X}_2$ of $\underset{\sim}{X}'\underset{\sim}{X}$

D. Compute $\underset{\sim}{C} = \underset{\sim}{X}'[\underset{\sim}{X}_2(\underset{\sim}{X}_2'\underset{\sim}{X}_2)^{-1}\underset{\sim}{X}_2' - \underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1']\,\underset{\sim}{X}$

E. Compute $\underset{\sim}{P}$ and $\underset{\sim}{D}$ as defined in section 2. The method discussed by Anderson (1958, pp. 340, 41) offers a convenient solution to finding $\underset{\sim}{P}$ and $\underset{\sim}{D}$. The following lemma, proved by Spjøtvoll (1972) may be useful.

Lemma Let q be the rank of $\underset{\sim}{X}$, q_o the number of predictors common in $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$ and $q_1 \leq q_2$ be the ranks of $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$, respectively. Then $q - q_1 - q_2 + 2q_o$ values of d_i are zero, $q_2 - q_1$ values of d_i are one, and $2(q_1 - q_o)$ are of the form $\pm d_i$, where $\{d_i\}$ are the diagonal elements of $\underset{\sim}{D}$.

F. Compute $\hat{\underset{\sim}{\gamma}} = \underset{\sim}{P}^{-1}\,\hat{\underset{\sim}{\beta}}$

G. Compute $\underset{\sim}{D}\,\hat{\underset{\sim}{\gamma}}$

H. Find the minimum and maximum roots, λ_{min} and λ_{max}, of the polynomial equation.

$$\sum_{i=1}^{q} d_i^2\,\hat{\gamma}_i^2(d_i - \lambda)^{-2} = qs^2F(1-\alpha:q,\ n-q)$$

Those roots are to be found in the following intervals. Let F be the set of indices i for which $\hat{\gamma}_i \neq 0$, then

$$\underset{i\in F}{MAX}\left\{d_i\right\} < \lambda_{max} \leq \underset{i\in F}{Max}\left\{d_i\right\} + \left[\frac{1}{c}\sum_{1}^{q} d_i^2\,\hat{\gamma}_i^2\right]^{1/2}$$

and $\underset{i\in F}{MIN}\left\{d_i\right\} - \left[\frac{1}{c}\sum_{1}^{q} d_i^2\,\hat{\gamma}_i^2\right]^{1/2} \leq \lambda_{min} < \underset{i\in F}{Min}\left\{d_i\right\}$

where $c = qs^2 \cdot F(1-\alpha:q,\ n-q)$

I. Let $a = -\lambda_{min}$ and $b = \lambda_{max}$. Determine the upper and lower limits of 100 $(1-\alpha)\%$ confidence interval $[A_1, A_2]$ as

$$A_1 = a\left(\sum_1^q \left[\frac{d_i \hat{\gamma}_i^2}{(a+d_i)}\right] - c\right)$$

and

$$A_2 = b\left(\sum_1^q \left[\frac{d_i \hat{\gamma}_i^2}{(b-d_i)}\right] + c\right)$$

J. If the interval $[A_1, A_2]$ does not contain the point zero, then reject the null hypothesis, $R_1^2 = R_2^2$, at the α level of significance.

5. Numerical Example

Scores on the predictor variables, transfer grade-point average, high school (per)centile rank, School College Ability Test: Total and the criterion, first semester grade-point average were available for a sample of eighty-one (81) transfer students who matriculated to a highly selective university in the midwest.

The multiple correlation involving the transfer grade-point average and high school rank will be compared with the multiple correlation involving transfer grade-point average and the SCAT score.

Let X be the 81 x 3 matrix of predictor variables. The columns of X are transfer gpa, high school (per)centile rank, and SCAT, respectively. The observations have been transformed to have a mean of zero and a variance of $1/(N-1)$.

The vector $\underset{\sim}{y}$ will be the 81 observed criterion scores, first semester gpa, and these data has also been transformed.

Since multiple correlations are invariant to linear transformations on X and $\underset{\sim}{y}$, the transformations do not affect the outcome and eases the computations because the intercept parameter has been removed.

The following computations correspond to the steps outlined in Section 4.

A. Not shown

B. $$X'X = \begin{bmatrix} 1.00000 & .55235 & .31485 \\ .55235 & 1.00000 & .52650 \\ .31485 & .52650 & 1.0000 \end{bmatrix}$$

$$\underset{\sim}{y}'X = (.53436, .42180, .30000)$$

C. i) $\hat{\beta} = (.43046, .13482, .09349)$

ii) $s^2 = .68505/75 = .00867$, $\alpha = .05$

iii) $$X_1'X_1 = \begin{bmatrix} 1.00000 & .55235 \\ .55325 & 1.00000 \end{bmatrix}$$

$$X_2'X_2 = \begin{bmatrix} 1.00000 & .31485 \\ .31485 & 1.00000 \end{bmatrix}$$

D. $$C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -.55691 & 0 \\ 0 & 0 & .72197 \end{bmatrix}$$

E. $$P = \begin{bmatrix} -.19129 & 1.00000 & -.63569 \\ -.30672 & 0.0 & 1.30443 \\ 1.14566 & 0.0 & -.26939 \end{bmatrix}$$

$$D = \begin{bmatrix} .89522 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -.89522 \end{bmatrix}$$

F. $\hat{\underset{\sim}{Y}}' = (.11210, .53436, .12971)$

G. $(\widehat{D\underset{\sim}{Y}})' = (.10036, 0.0, -.11612)$

H. $$\sum_{i=1}^{3} d_i^2 \hat{Y}_i^2 (d_i - \lambda)^2 = 3s^2 F(.95:3, 79)$$

becomes

$$-.07126\lambda^4 + .13778\lambda^2 - .0061\lambda - .02689 = 0$$

$\lambda_1 = .49945$, $\lambda_2 = -.44194$, $\lambda_3 = \lambda_{max} = 1.27902$ and $\lambda_4 = \lambda_{min} = -1.33649$

I. $A_1 = -.13412$ $\qquad A_2 = .11976$

J. Since the 95% simultaneous confidence interval, $[A_1, A_2]$ contains the point zero, we do not reject the null hypothesis of equal multiple correlations.

REFERENCES

Anderson, T.W. An Introduction to Multivariate Analysis. New York: Wiley, 1958.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the problem of nuisance parameters. Ann. Math. Statist. 11 271-283.

Spjøtvoll, E. (1972). Multiple comparison of regression functions. Ann. Math. Statist. 43 1076-1088.